CHALMERS
UNIVERSITY OF TECHNOLOGY

An Explicit-Implicit Splitting Method for a Convection-Diffusion Problem

Research Article

Vidar Thomée* and A. S. Vasudeva Murthy

An Explicit-Implicit Splitting Method for a Convection-Diffusion Problem

https://doi.org/10.1515/cmam-2018-0018
Received November 17, 2017; revised April 16, 2018; accepted June 7, 2018

Abstract: We analyze a second-order accurate finite difference method for a spatially periodic convection-diffusion problem. The method is a time stepping method based on the Strang splitting of the spatially semidiscrete solution, in which the diffusion part uses the Crank–Nicolson method and the convection part the explicit forward Euler approximation on a shorter time interval. When the diffusion coefficient is small, the forward Euler method may be used also for the diffusion term.

Keywords: Convection-Diffusion, Operator Splitting

MSC 2010: 35K10, 65M06, 65M15, 86A10

Dedicated to Amiya K. Pani on the occasion of his 60th birthday and part of the special issue "Recent Advances in PDE: Theory, Computations and Applications" in his honour.

1 Introduction

In this paper we shall consider the numerical solution of the following convection-diffusion problem in the cube $\Omega = (0, 2\pi)^d$:

$$\frac{\partial U}{\partial t} = \operatorname{div}(a\nabla U) + b \cdot \nabla U \quad \text{in } \Omega \quad \text{for } t > 0 \quad \text{with } U(0) = V, \tag{1.1}$$

under periodic boundary conditions, where the positive definite $d \times d$ matrix $a(x) = (a_{ij}(x))$ and the vector $b = b(x) = (b_1, \dots, b_d)$ are periodic and smooth.

Equation (1.1) is a special case of the initial-value problem for the operator equation

$$\frac{dU}{dt} = -\mathcal{A}U + \mathcal{B}U \quad \text{for } t \geq 0 \quad \text{with } U(0) = V, \tag{1.2}$$

where $\mathcal{A}$ and $\mathcal{B}$ represent different physical processes, in our case $\mathcal{A}U = -\operatorname{div}(a\nabla U)$, $\mathcal{B}U = b \cdot \nabla U$, with $\mathcal{A}$ representing a slow and $\mathcal{B}$ a fast physical process.

The solution of (1.2) may be formally expressed as

$$U(t) = \mathcal{E}(t)V = e^{-t(\mathcal{A}-\mathcal{B})}V \quad \text{for } t \geq 0.$$

To discretize such an equation in time, a common approach is to split $\mathcal{A} - \mathcal{B}$ into $\mathcal{A}$ and $-\mathcal{B}$. With k a time step one introduces $t_n = nk$ and one may then use the second-order symmetric Strang splitting [7, 8] on each time interval (t_{n-1}, t_n),

$$\mathcal{E}(k) = e^{-k(\mathcal{A}-\mathcal{B})} \approx e^{\frac{1}{2}k\mathcal{B}} e^{-k\mathcal{A}} e^{\frac{1}{2}k\mathcal{B}}, \tag{1.3}$$

which thus locally involves solutions of $U_t = -\mathcal{A}U$ and $U_t = \mathcal{B}U$, see e.g. Hundsdorfer and Verwer [5] and references therein.

***Corresponding author: Vidar Thomée,** Department of Mathematical Sciences, Chalmers University of Technology and University of Gothenburg, SE-412 96 Gothenburg, Sweden, e-mail: thomee@chalmers.se
A. S. Vasudeva Murthy, TIFR Centre for Applicable Mathematics, Yelahanka New Town, Bangalore 560 065, India, e-mail: vasu@math.tifrbng.res.in

The three exponentials on the right in (1.3) are then approximated by rational functions of $\mathcal{A}$ and $\mathcal{B}$, respectively, such as the Crank–Nicolson method $r_0(k\mathcal{A})$, with $r_0(\lambda) = (1 - \frac{1}{2}\lambda)/(1 + \frac{1}{2}\lambda)$, for the middle factor. The choice of approximation involving $\mathcal{B}$ is not so obvious. It has been suggested in the context of numerical weather prediction, e.g. in Baldauf [1], that time steps of different length could be used for the two processes, with shorter time intervals for the fast process and longer for the slow one. Also Gassmann and Herzog [3] discuss the difficulties associated with splitting in such situations. In the case of reaction-diffusion equation, see Estep, Ginting, Ropp, Shadid and Tavener [2] and references therein. Our aim in this work is to discuss this problem in a somewhat rigorous fashion for our simple model problem.

We note that if $\mathcal{A}$ and $\mathcal{B}$ commute, which holds for (1.1) when a and b are independent of x, then $e^{-k(\mathcal{A}-\mathcal{B})} = e^{-k\mathcal{A}}e^{k\mathcal{B}} = e^{\frac{1}{2}k\mathcal{B}}e^{-k\mathcal{A}}\,e^{\frac{1}{2}k\mathcal{B}}$ so that the error in (1.3) is zero. When $\mathcal{A}$ and $\mathcal{B}$ do not commute, then formally, by Taylor expansion, $e^{-k(\mathcal{A}-\mathcal{B})} - e^{-k\mathcal{A}}e^{k\mathcal{B}} = O(k^2)$ and $e^{-k(\mathcal{A}-\mathcal{B})} - e^{\frac{1}{2}k\mathcal{B}}e^{-k\mathcal{A}}e^{\frac{1}{2}k\mathcal{B}} = O(k^3)$. Error estimates for the time splitting, depending on the regularity of the initial values may be found in Jahnke and Lubich [6], Hansen and Ostermann [4] and references therein.

In the method that we study in this paper, we begin by discretizing (1.1) in the spatial variables. We let $h = \frac{2\pi}{M}$, where M is a positive integer, and define a corresponding uniform mesh

$$\Omega_h = \{x = x_j = jh : j = (j_1, \dots, j_d),\ 1 \le j_l \le M,\ l = 1, \dots, d\}. \tag{1.4}$$

For M-periodic vectors u with elements u_j, corresponding to the mesh-point x_j, and with $u_{j+Me_l} = u_j$, we consider the following simple second-order finite difference approximation of (1.1):

$$\frac{du}{dt} = \sum_{i,j=1}^{d} \bar\partial_j(\acute a_{ij}\partial_i u) + \sum_{j=1}^{d} \frac{1}{2} b_j(\partial_j + \bar\partial_j)u \quad \text{in } \Omega_h \quad \text{for } t > 0 \quad \text{with } u(0) = v. \tag{1.5}$$

Here ∂_j, $\bar\partial_j$ are forward and backward finite difference quotients in the direction of x_j, $\acute a_{ij}(x_l) = a_{ij}(x_l + \frac{1}{2}he_i)$, and v the restriction of V to Ω_h. Problem (1.5) may be written as a system of ODEs in time,

$$\frac{du}{dt} = -Au + Bu \quad \text{for } t > 0 \quad \text{with } u(0) = v, \tag{1.6}$$

where the $M^d \times M^d$ matrices A and B correspond to the differential operators $\mathcal{A}$ and $\mathcal{B}$. It is then to (1.6) that we will apply the splitting approach.

In the one-dimensional case we may take, with $h = \frac{2\pi}{M}$ the mesh-width and $x_l = lh$,

$$A = \frac{1}{h^2}\begin{bmatrix} d(x_1) & -a(x_{1.5}) & 0 & \dots & \dots & -a(x_{0.5}) \\ -a(x_{1.5}) & d(x_2) & -a(x_{2.5}) & \dots & \dots & 0 \\ 0 & -a(x_{2.5}) & d(x_3) & \dots & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \ddots & -a(x_{M-0.5}) \\ -a(x_{0.5}) & 0 & 0 & \dots & -a(x_{M-0.5}) & d(x_M) \end{bmatrix},$$

where $d(x_l) = a(x_l + 0.5h) + a(x_l - 0.5h)$ (recall $a(x_{M+0.5}) = a(x_{0.5})$), and

$$B = \frac{1}{2h}\begin{bmatrix} 0 & b(x_1) & \dots & \dots & -b(x_1) \\ -b(x_2) & 0 & b(x_2) & \dots & 0 \\ \vdots & \vdots & & \ddots & \\ \vdots & \vdots & & & b(x_{M-1}) \\ b(x_M) & 0 & \dots & -b(x_M) & 0 \end{bmatrix}.$$

The solution of (1.6), the spatially semidiscrete solution, is

$$u(t) = E(t)v = e^{-t(A-B)}v,$$

and we shall see that the error in this approximation is $O(h^2)$, under the appropriate regularity assumptions. For the time discretization we shall work with basic time intervals of length $k = \frac{1}{N}$, where $N = 2p$ is an even positive integer, then apply the Strang splitting (1.3) on each of the interval, so that

$$E(k) = e^{-k(A-B)} \approx e^{\frac{1}{2}kB}e^{-kA}e^{\frac{1}{2}kB}, \tag{1.7}$$

and finally approximate the three exponential factors.

We would like the time discretization error to match that of the discretization in space. For a second-order time discretization method, this will require $k = O(h)$. For $k \le \gamma h^2$, with γ appropriate, the problem may be solved by explicit approximations but since we prefer N to be relatively small, we will consider methods with h and k of the same order.

For the approximation of e^{-kA} in (1.7) we shall use the Crank–Nicolson method. Then, in order to approximate $e^{\frac{1}{2}kB}$, we would like to use the forward Euler method on a time interval of length $k_1 < k$. Assuming for the moment that b is constant and thus B skew-symmetric, we note that

$$\|I + k_1 B\| = (1 + k_1^2 \|B\|^2)^{\frac{1}{2}} \le 1 + Ck_1^2 h^{-2}.$$

Here and below, $\|\cdot\|$ denotes the standard matrix norm induced by the ℓ^2 vector inner product. Stability therefore holds if $k_1^2 h^{-2} \le Ck_1$, or if $k_1 \le Ch^2$. Since k should be of the same order as h, this makes it natural to choose $k_1 = k^2$. We thus subdivide the time intervals of length k into N subintervals of length $k^2 = \frac{k}{N}$ and apply an explicit forward Euler approximation on each of these. As we shall see, this approximation matches the second-order of the Crank–Nicolson method.

Thus the diffusion part of the equation is approximated on intervals of length k and the convection part on intervals of length k^2. We consider thus the time discrete solution at time $t_n = nk$,

$$u^n = E_k^n v, \quad \text{where } E_k = \mathcal{B}_k r_0(kA)\mathcal{B}_k \text{ with } \mathcal{B}_k = (I + k^2 B)^p. \tag{1.8}$$

In fact, in the successive time stepping, only the matrix $\tilde{\mathcal{B}}_k = \mathcal{B}_k^2 = (I + k^2 B)^N$ is used, except in the first and last steps. The method proposed thus replaces the solution at each time step of a nonsymmetric problem, by the solution of a symmetric problem, plus applications of an explicit method, but successively repeated $\frac{N}{2} = p$ times before and after the diffusion approximation, thus covering the interval of length $Nk^2 = k$. We re-emphasize that the splitting is applied to the spatially semidiscrete problem and not to the continuous problem.

The analysis sketched above is carried out in Section 2. The analysis will use discrete Sobolev norms. After this, in Section 2, we discuss the case when equation (1.1) contains a small diffusion coefficient ε. In this case we are able to show that if $\varepsilon \le \gamma h$ with γ sufficiently small, then the approximation of $e^{-k\varepsilon A}$ can be done by the forward Euler method, and, with the convection part as before, we have a purely explicit second-order approximation method. We close the paper by presenting some numerical illustrations in Section 4.

2 Basic Error Analysis

For the periodic problem in $\Omega = (0, 2\pi)^d$ and with Ω_h as in (1.4), we introduce the discrete inner product and norm

$$(v, w)_h = h^d \sum_{x_j \in \Omega_h} v_j w_j \quad \text{and} \quad \|v\|_h = (v, v)_h^{\frac{1}{2}}.$$

Further, we set $\partial_j u(x) = \frac{1}{h}(u(x + he_j) - u(x))$ and $\partial^\alpha u = \partial_1^{\alpha_1} \dots \partial_d^{\alpha_d} u$ for $\alpha = (\alpha_1, \dots, \alpha_d)$, and define the discrete Sobolev norm, with $|\alpha| = \alpha_1 + \cdots + \alpha_d$,

$$\|u\|_{h,s} = \Big(\sum_{|\alpha| \le s} \|\partial^\alpha u\|_h^2 \Big)^{\frac{1}{2}} \quad \text{for } s \ge 0.$$

We shall also use $\bar{\partial}_j u(x) = \frac{1}{h}(u(x) - u(x - he_j))$. We define

$$\|w\|_{\mathbb{C}^s(R)} = \max_{|\alpha| \le s} \sup_{x \in R} |D^\alpha w(x)|, \quad D^\alpha = \Big(\frac{\partial}{\partial x_1}\Big)^{\alpha_1} \cdots \Big(\frac{\partial}{\partial x_d}\Big)^{\alpha_d}.$$

We shall write $\mathbb{C}^s$ for $\mathbb{C}^s(\Omega)$, and $\mathbb{C}$ for $\mathbb{C}^0$. We note that defining U_h to be the restriction to the mesh Ω_h of a smooth function U, i.e. by $(U_h)_j = U(x_j)$, we have $\|U_h\|_{h,s} \le C\|U\|_{\mathbb{C}^s}$.

Consider now the matrices A and B in our convection-diffusion problem (1.2). They satisfy, with C independent of h,

$$\|Au\|_h \le C\|u\|_{h,2} \quad \text{and} \quad \|Bu\|_h \le C\|u\|_{h,1}. \tag{2.1}$$

Setting $Q(x) = \{y : |y_s - x_s| \le h,\ s = 1, \dots, d\}$, we have, since the terms in AU_h are symmetric difference quotients of U at the mesh-points x_j, and the terms in $(\mathcal{A}U)_h$ are the corresponding derivatives of U at x_j,

$$|(AU_h)(x_j) - (\mathcal{A}U)(x_j)| \le Ch^2 \|U\|_{\mathbb{C}^4(Q(x_j))} \tag{2.2}$$

and

$$|(BU_h)(x_j) - (\mathcal{B}U)(x_j)| \le Ch^2 \|U\|_{\mathbb{C}^3(Q(x_j))}, \tag{2.3}$$

expressing, in particular, that (1.5) is a second-order approximation of (1.1).

We note that, for $|\alpha| = s$,

$$\|\partial^\alpha Au - A\partial^\alpha u\|_h \le C\|u\|_{h,s+1} \quad \text{and} \quad \|\partial^\alpha Bu - B\partial^\alpha u\|_h \le C\|u\|_{h,s}, \tag{2.4}$$

and we may conclude from (2.1) that

$$\|Au\|_{h,s} \le C\|u\|_{h,s+2} \quad \text{and} \quad \|Bu\|_{h,s} \le C\|u\|_{h,s+1}. \tag{2.5}$$

The matrix A is positive semidefinite, with

$$\|v\|_{h,1}^2 \le C((Av, v)_h + \|v\|_h^2). \tag{2.6}$$

Further, it follows easily from the definition of A that

$$\|Av\|_h \le \alpha h^{-2}\|v\|_h, \quad \text{where } \alpha = 4d\lambda_{\max}(a). \tag{2.7}$$

For $B = (b_{il})$ we have

$$b_{il} = \begin{cases} \pm b_j(x_i) & \text{if } l = i \pm e_j,\ j = 1, \dots, d, \\ 0 & \text{for other } l, \end{cases}$$

and that then $b_{i+e_j,i} = -b_j(x_{i+e_j})$. It follows from this that

$$B = B_0 + B_1 \quad \text{with } B_0 = \frac{1}{2}(B - B^T),\ B_1 = \frac{1}{2}(B + B^T).$$

Here B_0 is skew-symmetric and

$$\|B_0\| \le h^{-1}\beta_0, \quad \beta_0 = \sum_{j=1}^d \|b_j\|_{\mathbb{C}}, \qquad \|B_1\| \le \beta_1 = \sum_{j=1}^d \left\| \frac{\partial b_j}{\partial x_j} \right\|_{\mathbb{C}}. \tag{2.8}$$

Note also that $(B_0 v, v) = 0$ for all v.

We begin with the straightforward standard analysis of the spatially semidiscrete problem (1.6), which we include for completeness. We first show the stability of the solution operator of (1.6) in discrete Sobolev norms.

Lemma 2.1. *Let $E(t) = e^{-t(A-B)}$. Then, for any $s \ge 0$, with $C = C_s$ independent of h,*

$$\|E(t)v\|_{h,s} \le e^{Ct}\|v\|_{h,s} \quad \textit{for } t \ge 0.$$

Proof. Let $s \ge 0$ and let $|\alpha| = s$. From (1.6) we find, for $u(t) = E(t)v$,

$$(\partial^\alpha u_t, \partial^\alpha u)_h + (\partial^\alpha Au, \partial^\alpha u)_h = (\partial^\alpha Bu, \partial^\alpha u)_h. \tag{2.9}$$

Here, by (2.4),

$$|(\partial^\alpha Au, \partial^\alpha u)_h - (A\partial^\alpha u, \partial^\alpha u)_h| \le C\|u\|_{h,s+1}\|u\|_{h,s}. \tag{2.10}$$

Further, since $B = B_0 + B_1$ with B_0 skew-symmetric,

$$|(\partial^\alpha Bu, \partial^\alpha u)_h| \le |(B\partial^\alpha u, \partial^\alpha u)_h| + C\|u\|_{h,s} \le C\|u\|_{h,s}^2.$$

Therefore, by (2.9),

$$\frac{1}{2}\frac{d}{dt}\|\partial^\alpha u\|_h^2 + (A\partial^\alpha u, \partial^\alpha u)_h + \|\partial^\alpha u\|_h^2 \le C\|u\|_{h,s+1}\|u\|_{h,s}.$$

Hence, using (2.6) and summing over $|\alpha| \le s$ we find, with $c > 0$,

$$\frac{d}{dt}\|u\|_{h,s}^2 + c\|u\|_{h,s+1}^2 \le C\|u\|_{h,s+1}\|u\|_{h,s} \le c\|u\|_{h,s+1}^2 + C\|u\|_{h,s}^2,$$

or, with $C = C_s$,

$$\frac{d}{dt}\|u\|_{h,s}^2 \le C\|u\|_{h,s}^2, \tag{2.11}$$

from which the lemma follows. □

Note that the special case of e^{-tA} is included for $B = 0$.

As a consequence, we have the following second-order error estimate.

Theorem 2.1. *We have, for the solutions of* (1.6) *and* (1.1), *with* $v = V_h$, *and* $C = C_T$,

$$\|u(t) - U_h(t)\|_h \le Ch^2\|V\|_{\mathbb{C}^4} \quad \text{for } nk \le T < \infty.$$

Proof. Setting $\omega = u - U_h$, we find

$$\frac{d\omega}{dt} = -A\omega + B\omega + \rho \quad \text{in } \Omega_h \quad \text{for } t > 0 \quad \text{with } \omega(0) = 0,$$

where $\rho = ((\mathcal{A}U)_h - AU_h) - ((\mathcal{B}U_h) - BU_h)$. Here, by (2.2) and (2.3),

$$\|\rho(t)\|_h \le \|((\mathcal{A}U)_h - AU_h)(t)\|_h + \|((\mathcal{B}U)_h - BU_h)(t)\|_h \le Ch^2\|U(t)\|_{\mathbb{C}^4},$$

and hence

$$\omega(t) = \int_0^t E(t-s)\rho(s)\,ds,$$

where $\|\rho(s)\|_h \le Ch^2\|U(s)\|_{\mathbb{C}^4}$. Hence, by Lemma 2.1, since $\|U(s)\|_{\mathbb{C}^4} \le C\|V\|_{\mathbb{C}^4}$,

$$\|\omega(t)\|_h \le C\int_0^t \|\rho(s)\|_h\,ds \le Ch^2\int_0^t \|U(s)\|_{\mathbb{C}^4}\,ds \le Ch^2\|V\|_{\mathbb{C}^4}.$$ □

Turning to the analysis of the time discretization, we first show the stability of e^{tB}.

Lemma 2.2. *For any* $s \ge 0$, *we have, with* C_s *independent of* h,

$$\|e^{tB}v\|_{h,s} \le e^{C_s t}\|v\|_{h,s} \quad \text{for } t \ge 0. \tag{2.12}$$

Here we may choose $C_0 = \beta_1$, *as in* (2.8).

Proof. Let $s \ge 0$ and let $|\alpha| = s$. Then for the solution of (1.6) with $A = 0$,

$$(\partial^\alpha u_t, \partial^\alpha u)_h = (\partial^\alpha Bu, \partial^\alpha u)_h = (B\partial^\alpha u, \partial^\alpha u)_h + Q_\alpha,$$

where $|Q_\alpha| \le C_s\|u\|_{h,s}^2$. Since $(Bu, u)_h = (B_1 u, u)_h \le \beta_1\|u\|_h^2$, we conclude that (2.11) holds, which shows (2.12). For $s = 0$ we have $Q_\alpha = 0$ and hence (2.11) holds with $C = 2\beta_1$ which implies (2.12), with $C_0 = \beta_1$. □

We now show the following error estimate for the Strang splitting.

Lemma 2.3. *We have, with* C *independent of* h *and* k,

$$\|e^{-k(A-B)}v - e^{\frac{1}{2}kB}e^{-kA}\,e^{\frac{1}{2}kB}v\|_h \le Ck^3\|v\|_{h,6}.$$

Proof. Setting $F(k) = e^{-k(A-B)} - e^{\frac{1}{2}kB}e^{-kA}\,e^{\frac{1}{2}kB}$ and noting that $F(0) = F'(0) = F''(0) = 0$, we may use Taylor's formula to obtain

$$\|F(k)\| = \left\|(F(k) - F(0) - kF'(0) - \frac{1}{2}k^2F''(0))v\right\|_h \le \frac{1}{6}k^3\sup_{s\le k}\|F'''(s)v\|_h.$$

Here, for $s \le k$, using (2.1) and Lemmas 2.1 and 2.2,

$$\|F'''(s)v\|_h \le \|(A-B)^3 e^{-k(A-B)}v\|_h + C \sum_{i_1+i_2+i_3=3} \|B^{i_1} e^{\frac{1}{2}sB} A^{i_2} e^{-sA} B^{i_3} e^{\frac{1}{2}sB} v\|_h$$

$$\le C\Big(\|v\|_{h,6} + \sum_{i_1+i_2+i_3=3} \|v\|_{h,i_1+2i_2+i_3}\Big) \le C\|v\|_{h,6},$$

which shows the lemma. □

We now turn to the time stepping operator E_k defined in (1.8) and begin with the following stability result.

Lemma 2.4. *Let $k \le \gamma h$. Then, with β_0, β_1 as in* (2.8),

$$\|\mathcal{B}_k\| = \|(I + k^2 B)^p\| \le e^{\frac{1}{2}\mu k}, \quad \text{where } \mu = \frac{1}{2}(\gamma\beta_0)^2 + \beta_1. \tag{2.13}$$

Further, $E_k = \mathcal{B}_k r_0(kA)\mathcal{B}_k$ is stable, and

$$\|E_k^n\| \le e^{\mu T} \quad \text{for } nk \le T. \tag{2.14}$$

Proof. Since B_0 is skew-symmetric, we have

$$\|I + k^2 B_0\|^2 = 1 + k^4\|B_0\|^2 \le 1 + k^4 h^{-2}\beta_0^2 \le 1 + (\gamma\beta_0)^2 k^2 \le e^{(\gamma\beta_0)^2 k^2}.$$

Hence

$$\|I + k^2 B\| \le \|I + k^2 B_0\| + k^2\|B_1\| \le e^{\frac{1}{2}(\gamma\beta_0)^2 k^2} + \beta_1 k^2 \le e^{\mu k^2}, \tag{2.15}$$

which shows (2.13) since $2pk = 1$. Hence (2.14) follows by $\|r_0(kA)\| \le 1$. □

We start the analysis of the time discretization error with the following.

Lemma 2.5. *Let M be a square matrix, and assume $\|e^{sM}\| \le C$ for $s \le t$. Then we have, for $s \le t$,*

$$\|(e^{tM} - (I + tM))v\|_h \le Ct^2\|M^2 v\|_h$$

and

$$\|(e^{tM} - (I + tM + \frac{1}{2}t^2M^2))v\|_h \le Ct^3\|M^3 v\|_h. \tag{2.16}$$

If also $\|(I + \frac{1}{2}sM)^{-1}\| \le C$ for $s \le t$, then

$$\|(e^{-tM} - r_0(tM))v\|_h \le Ct^3(\|M^3 v\|_h + \|v\|_h). \tag{2.17}$$

Proof. By Taylor expansion we have

$$e^{tM} = I + tM + \int_0^t (t-s)e^{Ms}M^2\,ds,$$

and hence

$$\|(e^{tM} - (I + tM))v\|_h \le C\int_0^t (t-s)\|M^2 v\|_h\,ds = \frac{1}{2}Ct^2\|M^2 v\|_h.$$

Estimate (2.16) follows analogously. For (2.17), we use (2.16) together with

$$r_0(tM) = I - tM + \frac{1}{2}t^2M^2 + \frac{1}{2}\int_0^t (t-s)^2 r_0'''(sM)\,ds,$$

where $r_0'''(\lambda) = -\frac{3}{2}(1 + \frac{1}{2}\lambda)^{-4}$, to complete the proof. □

We now show the following error estimate for $\mathcal{B}_k$.

Lemma 2.6. *Let $k \le \gamma h$. Then we have, with $C = C_\gamma$,*

$$\|(e^{\frac{1}{2}kB} - \mathcal{B}_k)v\|_h \le Ck^3\|v\|_{h,2}.$$

Proof. Since $pk^2 = \frac{1}{2}k$, we may write

$$e^{\frac{1}{2}kB}v - (1 + k^2 B)^p v = \sum_{j=0}^{p-1} e^{(p-j-1)k^2 B}(e^{k^2 B} - (I + k^2 B))(I + k^2 B)^j v.$$

By Lemma 2.5, we have

$$\|(e^{k^2B} - (1 + k^2B))v\|_h \le Ck^4\|B^2v\|_h.$$

By Lemma 2.2, $\|e^{(p-j-1)k^2B}\| \le e^{\frac{1}{2}k\beta_1}$ for $j \le p-1$. Using also (2.15), we find

$$\begin{aligned}\|e^{\frac{1}{2}kB}v - (1 + k^2B)^p v\|_h &\le Ck^4 \sum_{j=0}^{p-1} \|B^2(I + k^2B)^j v\|_h \\ &\le Ck^4 \sum_{j=0}^{p-1} e^{\mu jk^2}\|B^2v\|_h \le Ck^4 p e^{\frac{1}{2}\mu k}\|B^2v\|_h \le Ck^3\|v\|_{h,2}\end{aligned}$$

which completes the proof. □

We now show the following error estimate for E_k.

Lemma 2.7. *Let $k \le \gamma h$. Then we have, with $C = C_\gamma$,*

$$\|E(k)v - E_k v\|_h \le Ck^3\|v\|_{h,6}.$$

Proof. In view of Lemma 2.3, it remains to show

$$\|e^{\frac{1}{2}kB}e^{-kA}e^{\frac{1}{2}kB}v - E_k v\|_h \le Ck^3\|v\|_{h,6}.$$

We have

$$\begin{aligned}e^{\frac{1}{2}kB}e^{-kA}e^{\frac{1}{2}kB} - \mathcal{B}_k r_0(kA)\mathcal{B}_k &= (e^{\frac{1}{2}kB} - \mathcal{B}_k)e^{-kA}e^{\frac{1}{2}kB} + \mathcal{B}_k(e^{-kA} - r_0(kA))e^{\frac{1}{2}kB} + \mathcal{B}_k r_0(kA)(e^{\frac{1}{2}kB} - \mathcal{B}_k) \\ &= J_1 + J_2 + J_3.\end{aligned}$$

Here, by the above lemmas,

$$\begin{aligned}&\|J_1 v\|_h \le Ck^3\|e^{-kA}e^{\frac{1}{2}kB}v\|_{h,2} \le Ck^3\|v\|_{h,2}, \\ &\|J_2 v\|_h \le Ck^3(\|A^3 e^{\frac{1}{2}kB}v\|_h + \|e^{\frac{1}{2}kB}v\|_h) \le Ck^3\|e^{\frac{1}{2}kB}v\|_{h,6} \le Ck^3\|v\|_{h,6}, \\ &\|J_3 v\|_h \le Ck^3\|v\|_{h,2},\end{aligned}$$

which completes the proof. □

We can now prove the following error estimate:

Theorem 2.2. *Let $k \le \gamma h$. Then we have for the solutions of* (1.8) *and* (1.6)*, with $C = C_{\gamma,T}$,*

$$\|u^n - u(nk)\|_h \le Ck^2\|v\|_{h,6} \quad \text{for } nk \le T.$$

Proof. We write

$$u^n - u(nk) = E_k^n v - E(nk)v = \sum_{j=0}^{n-1} E_k^{n-1-j}(E_k - E(k))E(jk). \tag{2.18}$$

Using Lemmas 2.4, 2.7 and 2.1, we obtain, for $nk \le T$,

$$\begin{aligned}\|E_k^n v - E(nk)v\|_h &\le C\sum_{j=0}^{n-1} \|(E_k - E(k))E(jk)v\|_h \\ &\le Ck^3 \sum_{j=0}^{n-1} \|E(jk)v\|_{h,6} \le Cnk^3\|v\|_{h,6} \le Ck^2\|v\|_{h,6}.\end{aligned}$$

□

Since $\|V_h\|_{h,6} \le C\|V\|_{\mathbb{C}^6}$, we immediately obtain from Theorems 2.1 and 2.2 the following total error estimate.

Theorem 2.3. *Let $v = V_h$ and $k \le \gamma h$. Then we have for the solutions of* (1.8) *and* (1.1)*, with $C = C_{\gamma,T}$,*

$$\|u^n - U_h(nk)\|_h \le Ch^2\|V\|_{\mathbb{C}^6} \quad \text{for } nk \le T.$$

3 The Case of a Small Diffusion Coefficient

In this section, we consider the variant of problem (1.1) with a small diffusion coefficient $\varepsilon > 0$,

$$\frac{\partial U}{\partial t} = \varepsilon \operatorname{div}(a\nabla U) + b \cdot \nabla U \quad \text{in } \Omega \quad \text{for } t > 0 \quad \text{with } U(0) = V.$$

The corresponding semidiscrete system (1.6) may then be written

$$\frac{du}{dt} = -\varepsilon Au + Bu \quad \text{for } t > 0 \quad \text{with } u(0) = v, \tag{3.1}$$

where A and B are as before. We shall see that (3.1) is stable, and satisfies an $O(h^2)$ error estimate, independently of ε. Further, for ε and k small, or more precisely, if $\max(\varepsilon, k) \le \gamma h$ and $k\varepsilon \le \frac{2h^2}{\alpha}$, we will be able to show an $O(k^2)$ estimate for the time discretization error, even when we use the less accurate forward Euler method for the A part of the time stepping operator, and with weaker regularity requirements than earlier. Also, we do not need to use the symmetric Strang splitting, and consider now, with $r_1(\lambda) = 1 - \lambda$,

$$U^n = \widetilde{E}_k^n v, \quad \text{where } \widetilde{E}_k = r_1(k\varepsilon A)\tilde{\mathcal{B}}_k \text{ with } \tilde{\mathcal{B}}_k = \mathcal{B}_k^2 = (I + k^2 B)^{2p}.$$

We note the inverse inequality $h\|u\|_{h,s} \le C\|u\|_{h,s-1}$, and hence

$$\varepsilon\|Au\|_{h,s} \le C\|u\|_{h,s+1} \quad \text{for } \varepsilon \le \gamma h \quad \text{if } \gamma > 0. \tag{3.2}$$

As in Section 2, we first attend to the spatially semidiscrete problem.

Lemma 3.1. *Let $\widetilde{E}(t) = e^{-t(\varepsilon A - B)}$. Then, for any $s \ge 0$, we have, with $C = C_s$ independent of $\varepsilon > 0$ and $h > 0$,*

$$\|\widetilde{E}(t)v\|_{h,s} \le e^{Ct}\|v\|_{h,s} \quad \text{for } t \ge 0.$$

Proof. Following the steps in the proof of Lemma 2.1, we have, for $u(t) = \widetilde{E}(t)v$ and $|\alpha| = s$,

$$(\partial^\alpha u_t, \partial^\alpha u)_h + \varepsilon(\partial^\alpha Au, \partial^\alpha u)_h = (\partial^\alpha Bu, \partial^\alpha u)_h. \tag{3.3}$$

Here, as in the proof of Lemma 2.1, $|(\partial^\alpha Bu, \partial^\alpha u)_h| \le C\|u\|_{h,s}^2$. Hence, by (3.3) and using (2.10),

$$\frac{1}{2}\frac{d}{dt}\|\partial^\alpha u\|_h^2 + \varepsilon(A\partial^\alpha u, \partial^\alpha u)_h + \varepsilon\|\partial^\alpha u\|_h^2 \le C\varepsilon\|u\|_{h,s+1}\|u\|_{h,s} + C\|u\|_{h,s}^2,$$

and thus, by (2.6), with $c > 0$,

$$\frac{d}{dt}\|u\|_{h,s}^2 + \varepsilon c\|u\|_{h,s+1}^2 \le \varepsilon c\|u\|_{h,s+1}^2 + C\|u\|_{h,s}^2.$$

This implies (2.11), with C independent of ε and h, and thus completes the proof. □

In the same way as in Section 2, the stability shows the following error estimate.

Theorem 3.1. *We have, for the solutions of* (3.1) *and* (1.1)*, with $C = C_T$ independent of ε,*

$$\|u(t) - U_h(t)\|_h \le Ch^2\|V\|_{\mathbb{C}^4} \quad \text{for } nk \le T.$$

In the analysis of the time discretization we begin with the analogue of Lemma 2.3.

Lemma 3.2. *We have, with $C = C_\gamma$ independent of h and k,*

$$\|E(k)v - e^{-k\varepsilon A}\, e^{kB}v\|_h \le C(\varepsilon k^2 + k^3)\|v\|_{h,3} \quad \text{for } \varepsilon \le \gamma h.$$

Proof. With $F(k) = e^{-k(\varepsilon A - B)} - e^{-k\varepsilon A}\, e^{kB}$, we have $F(0) = F'(0) = 0$ and hence, by Taylor's formula,

$$\|F(k)v\|_h = \|(F(k) - F(0) - kF'(0))v\|_h \le \frac{1}{2}k^2 \sup_{s\le k}\|F''(s)v\|_h.$$

Here

$$\begin{aligned} F''(s) &= e^{-s(\varepsilon A-B)}(\varepsilon A-B)^2 - e^{-s\varepsilon A}(\varepsilon^2A^2 - 2\varepsilon AB + B^2)e^{sB} \\ &= e^{-s(\varepsilon A-B)}(\varepsilon^2A^2 - \varepsilon AB - \varepsilon BA) - e^{-s\varepsilon A}(\varepsilon^2A^2 - 2\varepsilon AB)e^{sB} + (e^{-s(\varepsilon A-B)} - e^{-s\varepsilon A}e^{sB})B^2 \\ &= G_1(s) + G_2(s) + G_3(s). \end{aligned}$$

Using (3.2), (2.5) and the boundedness of the exponentials, we find

$$\|G_1(s)v + G_2(s)v\|_h \le C\varepsilon\|v\|_{h,3} \quad \text{for } s \le k.$$

Further, $G_3(0) = 0$ and hence

$$\|G_3(s)v\|_h \le s\sup_{\sigma\le s}\|G_3'(\sigma)v\|_h \le Ck\|v\|_{h,3} \quad \text{for } s \le k.$$

Together these estimates complete the proof of the lemma. □

We now turn to the time stepping operator $\widetilde{E}_k$ and begin with its stability.

Lemma 3.3. *If $k\varepsilon \le \frac{2h^2}{\alpha}$, then*

$$\|r_1(k\varepsilon A)\| = \|I - k\varepsilon A\| \le 1. \tag{3.4}$$

If also $k \le \gamma h$, then $\widetilde{E}_k$ is stable, or, with μ as in Lemma 2.4,

$$\|\widetilde{E}_k^n\| \le e^{\mu T} \quad \textit{for } nk \le T.$$

Proof. We note that, since A is positive semidefinite,

$$\|I - k\varepsilon A\| \le 1 \quad \text{if } k\varepsilon\|A\| \le 2,$$

and thus (3.4) holds by (2.7). Hence, by Lemma 2.4,

$$\|\widetilde{E}_k\| \le \|r_1(kA)\|\|\mathcal{B}_k\|^2 \le e^{\mu k}.$$ □

We now turn to the error analysis and show the following.

Lemma 3.4. *If $\max(\varepsilon, k) \le \gamma h$ and $k\varepsilon \le \frac{2h^2}{\alpha}$, we have*

$$\|E(k)v - \widetilde{E}_k v\|_h \le C(\varepsilon k^2 + k^3)\|v\|_{h,3}.$$

Proof. In view of Lemma 3.2 it remains to show

$$\|e^{-k\varepsilon A}e^{kB}v - \widetilde{E}_k v\|_h \le C(\varepsilon k^2 + k^3)\|v\|_{h,3}.$$

We first note that by Lemma 2.5,

$$\|(e^{-k\varepsilon A} - r_1(k\varepsilon A))v\|_h \le C\varepsilon^2k^2\|A^2v\|_h \le C\varepsilon k^2\|v\|_{h,3}. \tag{3.5}$$

We write

$$e^{-k\varepsilon A}e^{kB} - \widetilde{E}_k = (e^{-k\varepsilon A} - r_1(k\varepsilon A))e^{kB} + r_1(k\varepsilon A)(e^{kB} - \tilde{\mathcal{B}}_k) = J_1 + J_2.$$

Here by (3.5) and Lemma 2.5, and by (3.4) and Lemma 2.4,

$$\|J_1v\|_h \le C\varepsilon k^2\|e^{kB}v\|_{h,3} \le C\varepsilon k^2\|v\|_{h,3} \quad \text{and} \quad \|J_2v\|_h \le Ck^3\|v\|_{h,2},$$

which completes the proof □

The following is the resulting error estimate.

Theorem 3.2. *If $\max(\varepsilon, k) \le \gamma h$ and $k\varepsilon \le 2h^2/\alpha$ we have, with $C = C_{\gamma,T}$ independent of h, k and ε,*

$$\|u^n - u(nk)\|_h \le C(\varepsilon k + k^2)\|v\|_{h,3} \quad \textit{for } nk \le T.$$

Proof. Using the analogue of (2.18), we find

$$\|\widetilde{E}_k^n v - E(nk)v\|_h \le C\sum_{j=0}^{n-1}\|(\widetilde{E}_k - E(k))E(jk)v\|_h \le C(\varepsilon k^2 + k^3)\sum_{j=0}^{n-1}\|E(jk)v\|_{h,3} \le CT(\varepsilon k + k^2)\|v\|_{h,3}.$$ □

As in Section 2, our error estimates in Theorems 3.1 and 3.2 together show a total error estimate.

Theorem 3.3. *With $v = V_h$ and for $\max(\varepsilon, k) \le \gamma h$ and $k\varepsilon \le \frac{2h^2}{\alpha}$ we have, with $C = C_{\gamma,T}$ independent of h, k and ε,*

$$\|u^n - U_h(nk)\|_h \le Ch^2 \|V\|_{\mathbb{C}^4} \quad \text{for } nk \le T.$$

4 Numerical Illustrations

In this section we present some numerical computations to illustrate our error estimates. We restrict ourselves to the one-dimensional version of (1.1),

$$U_t = (a\,U_x)_x + b\,U_x \quad \text{for } x \in (0, 2\pi), \quad t > 0, \quad \text{with } U(x, 0) = \sin x. \tag{4.1}$$

As before we shall choose $h = \frac{2\pi}{M}$, $k = \frac{1}{N}$, where M and N are positive integers, and study the effect of doubling these integers.

We begin with the simple case $a = b = 1$, in which case the exact solution is $U(x, t) = e^{-t}\sin(x + t)$. In Table 1 we compile the errors in the numerical solution at $t = 1$, first in the spatial discretization, then when our time stepping method is applied to the semidiscrete solution, and finally the total error. We use $N = 4, 8, 16, 32, 64$, and $M = 5N$, so that $\frac{k}{h} = \frac{5}{2\pi} = 0.7958$. The successive ratios of the total errors are given in the last column and confirm the second-order convergence estimates resulting from Theorems 2.1–2.3.

We recall that in the case that a and b are constant, the matrices A and B involved in our method commute, and consequently the splitting error given in Lemma 2.3 vanishes. In order to also consider a situation when this does not happen, we let $a(x) = 1 + \frac{1}{2}\cos x$ and $b(x) = 1 + \frac{1}{2}\sin x$. To indicate that the matrices A and B do not commute in this case, we consider the corresponding continuous operators

$$\mathcal{A}U = -\left(\left(1 + \frac{1}{2}\cos x\right)U_x\right)_x, \quad \mathcal{B}U = \left(1 + \frac{1}{2}\sin x\right)U_x,$$

and find, after some effort,

$$\begin{aligned}(\mathcal{A}\mathcal{B} - \mathcal{B}\mathcal{A})U &= -(a(bU_x)_x)_x + b(aU_x)_{xx} \\ &= \left(1 - \frac{1}{4}\frac{2 - \sin x}{2 + \cos x}\right)U_x - \left(\frac{1}{2} + \frac{1}{2}\sin x - \frac{1}{4}\sin^2 x + \cos x\right)U_{xx}.\end{aligned}$$

Thus $\mathcal{A}$ and $\mathcal{B}$ do not commute, and therefore neither could A and B. The exact solution U is taken to be the semidiscrete solution with $M = 2560$, $N = 512$. The errors are presented in Table 2. Again we see that the errors are of second order, which agrees with the error bounds in Theorems 2.1–2.3.

We finally consider a numerical example for Section 3, for which we use (4.1) with $a = \varepsilon = 0.01$, $b = 1$. Here $U(x, t) = e^{-\varepsilon t}\sin(x + t)$, and $u^n = \tilde{E}_k^n v$ with $\tilde{E}_k = r_1(kA)\tilde{\mathcal{B}}_k$. Note that the condition $\varepsilon k \le \frac{2h^2}{\alpha}$, with $\alpha = 4$, now reduces to $\varepsilon \le \frac{\pi}{5}h < 0.8h$, or $\varepsilon \le \frac{\pi^2}{800} = 0.0123$ for $N = 64$, which is satisfied for our choice of ε. The results are given in Table 3 and agree with the error bounds of Section 3.

M	N	$\|(u - U_h)(\cdot, 1)\|_h$	$\|u^N - u(\cdot, 1)\|_h$	$\|u^N - U_h(\cdot, 1)\|_h$	Ratio
20	4	0.01670	0.01199	0.02494	
40	8	0.00423	0.00300	0.00621	4.01
80	16	0.00106	0.00075	0.00155	4.01
160	32	0.00027	0.00019	0.00039	3.97
320	64	0.00007	0.00005	0.00010	3.90

Table 1: Numerical results for constant coefficients. Here $h = \frac{2\pi}{M}$, $k = \frac{1}{N}$.

M	N	$\|(u-U_h)(\cdot,1)\|_h$	$\|u^N-u(\cdot,1)\|_h$	$\|u^N-U_h(\cdot,1)\|_h$	Ratio
20	4	0.02641	0.01419	0.03323	
40	8	0.00651	0.00356	0.00817	4.07
80	16	0.00163	0.00089	0.00203	4.02
160	32	0.00041	0.00022	0.00051	3.98
320	64	0.00010	0.00005	0.00013	3.92

Table 2: Numerical results for variable coefficients. Here $h=\frac{2\pi}{M}$, $k=\frac{1}{N}$.

M	N	$\|(u-U_h)(\cdot,1)\|_h$	$\|u^N-u(\cdot,1)\|_h$	$\|u^N-U_h(\cdot,1)\|_h$	Ratio
20	4	0.02872	0.05381	0.06237	
40	8	0.00721	0.01365	0.01555	4.01
80	16	0.00180	0.00342	0.00388	4.00
160	32	0.00045	0.00085	0.00097	4.00
320	64	0.00011	0.00021	0.00024	4.04

Table 3: Numerical results for constant coefficients and with small diffusion. Here $a=0.01$, $b=1$, $h=\frac{2\pi}{M}$, $k=\frac{1}{N}$.

References

[1] M. Baldauf, Linear stability analysis of Runge–Kutta-based partial time-splitting schemes for the Euler equations, *Monthly Weather Rev.* **138** (2010), 4475–4496.

[2] D. Estep, V. Ginting, D. Ropp, J. N. Shadid and S. Tavener, An a posteriori-a priori analysis of multiscale operator splitting, *SIAM J. Numer. Anal.* **46** (2008), no. 3, 1116–1146.

[3] A. Gassmann and H.-J. Herzog, A consistent time-split numerical scheme applied to the nonhydrostatic compressible equations, *Monthly Weather Rev.* **135** (2007), 20–36.

[4] E. Hansen and A. Ostermann, Exponential splitting for unbounded operators, *Math. Comp.* **78** (2009), no. 267, 1485–1496.

[5] W. Hundsdorfer and J. Verwer, *Numerical Solution of Time-Dependent Advection-Diffusion-Reaction Equations*, Springer Ser. Comput. Math. 33, Springer, Berlin, 2003.

[6] T. Jahnke and C. Lubich, Error bounds for exponential operator splittings, *BIT* **40** (2000), no. 4, 735–744.

[7] S. MacNamara and G. Strang, Operator splitting, in: *Splitting Methods in Communication, Imaging, Science, and Engineering*, Sci. Comput., Springer, Cham (2016), 95–114.

[8] G. Strang, On the construction and comparison of difference schemes, *SIAM J. Numer. Anal.* **5** (1968), 506–517.